# RNA Standards for Microarray Assays

NIST RNA Standards Meeting

28 March 2003

# General Principles

- The question of standards needs to be addressed for each species that will be profiled.

- At present, the primary targets for standard development would be human, mouse, and rat because of their importance for medical research.

- Standards developed should be based on principles that are extensible to other species.

- For the purposes of this workshop, we will focus on defining approaches to developing an optimal human standard.

# What is an RNA Standard?

- There are at least *four* standards that one might want to consider:
  - A common reference RNA to which all samples are compared
  - A common reference RNA that can be used to validate the measured expression level of particular genes on a particular platform
  - A common reference RNA which can be used to measure the performance of each platform
  - A common reference RNA that can be used as a “spiking” control to facilitate comparisons

# A Good Standard Should

- Allow performance validation of any single platform over time.
- Facilitate comparison between various platforms used to assay gene expression.
- Be constructed in such a manner as to assure consistency over time.
- Include a well defined protocol describing how it is made and validated.
- Include two or more samples that allow one to make both absolute and relative measurements of the abundance of individual transcripts.
- Not be limited to hybridization-based approaches, but should be amenable to use with other assays such as QRT-PCR

# Implementing a Standard

- An RNA standard would likely consist of two or more validated and qualified complex pools of RNA with some number of individual transcripts represented at some fixed absolute/relative concentrations.

- It was envisioned that NIST would maintain a primary standard and qualify secondary standards.

# Questions for this Workshop

- How much of the genome should be covered by any standard?

- What performance metrics should be established for any standard?

- What mix of “real” RNA and synthetic/exogenous species should be included in any standard? What should the source of the standard be?

- What metrics would be used for validation of the standard?

- How the standard would be used for calibration and validation of various platforms?

# Other views

- **Matthew Marton, Rosetta Inpharmatics**

  ***Properties and Uses of RNA Reference Standards in a Breast Cancer Clinical Study***

- **Paul Wolber, Agilent Technologies**

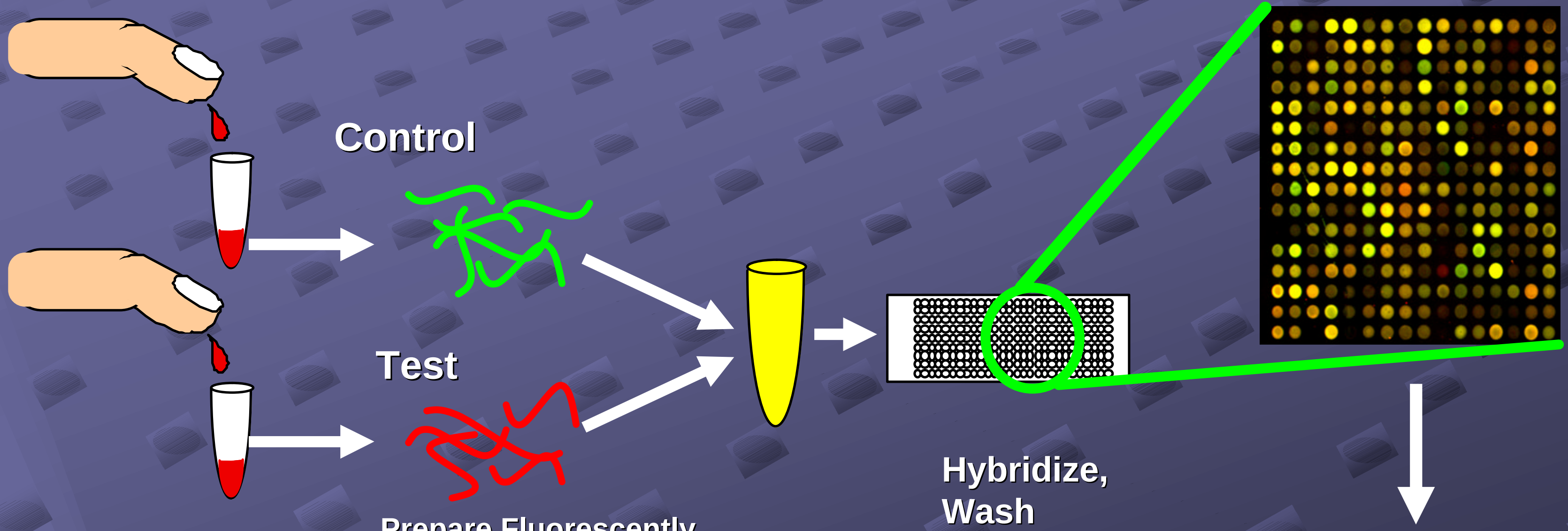
  ***Scale-up of an RNA Reference Standard for High-Throughput Microarray QC***

- **Natalia Novoradosky, Stratagene**

  ***Universal Reference RNA as a Standard for Microarray Expression Experiments***

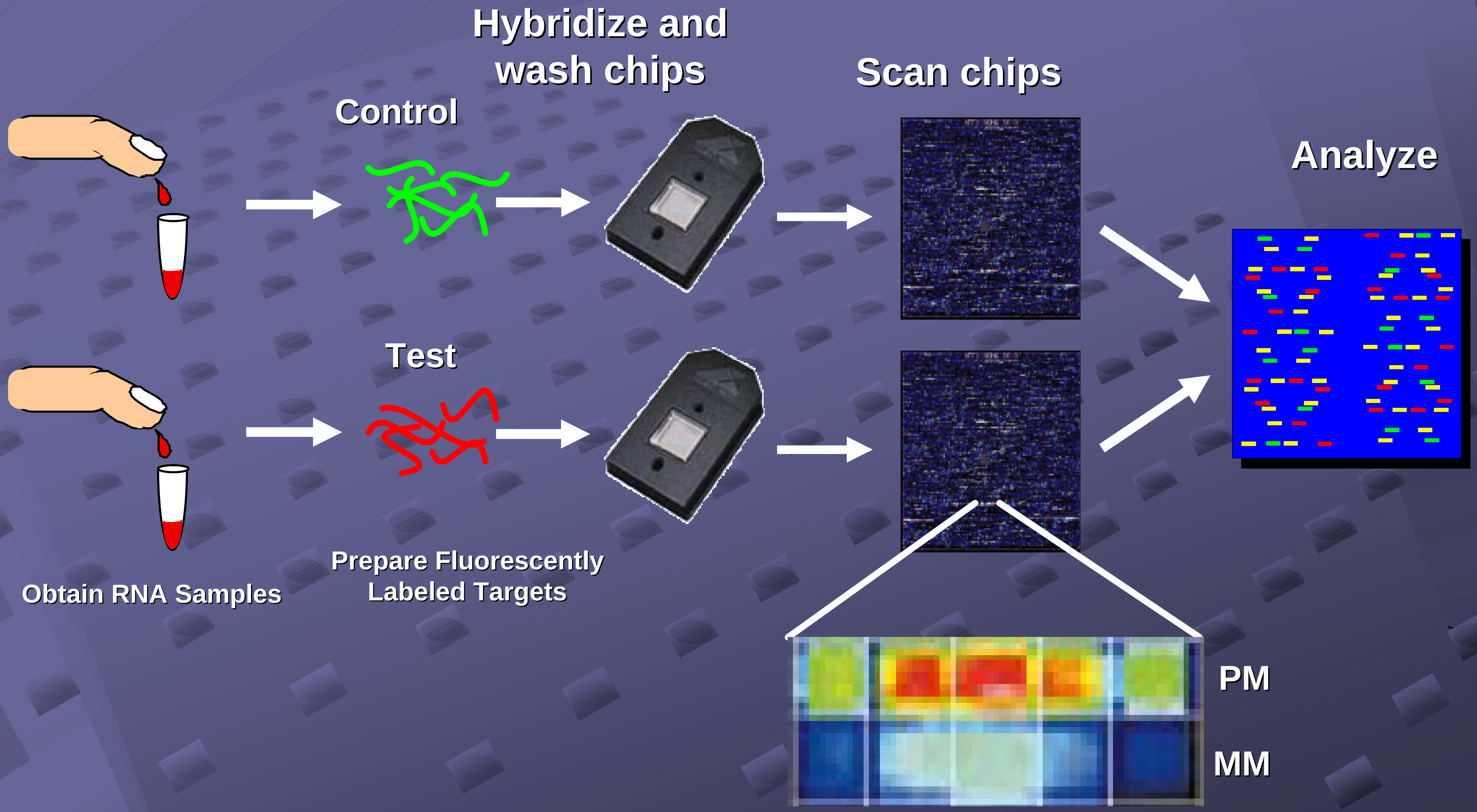
# We need a Universal Standard for validating Platform Performance

# Microarray Gene Chip Overview


Measure Fluoresence in 2 channels red/green
Control
Test
Hybridize, Wash
Obtain RNA Samples
Prepare Fluorescently Labeled Targets

Analyze the data to identify “differentially” expressed genes

# Affymetrix GeneChip™ Expression Analysis



Hybridize and wash chips

Scan chips

Control

Analyze

Test

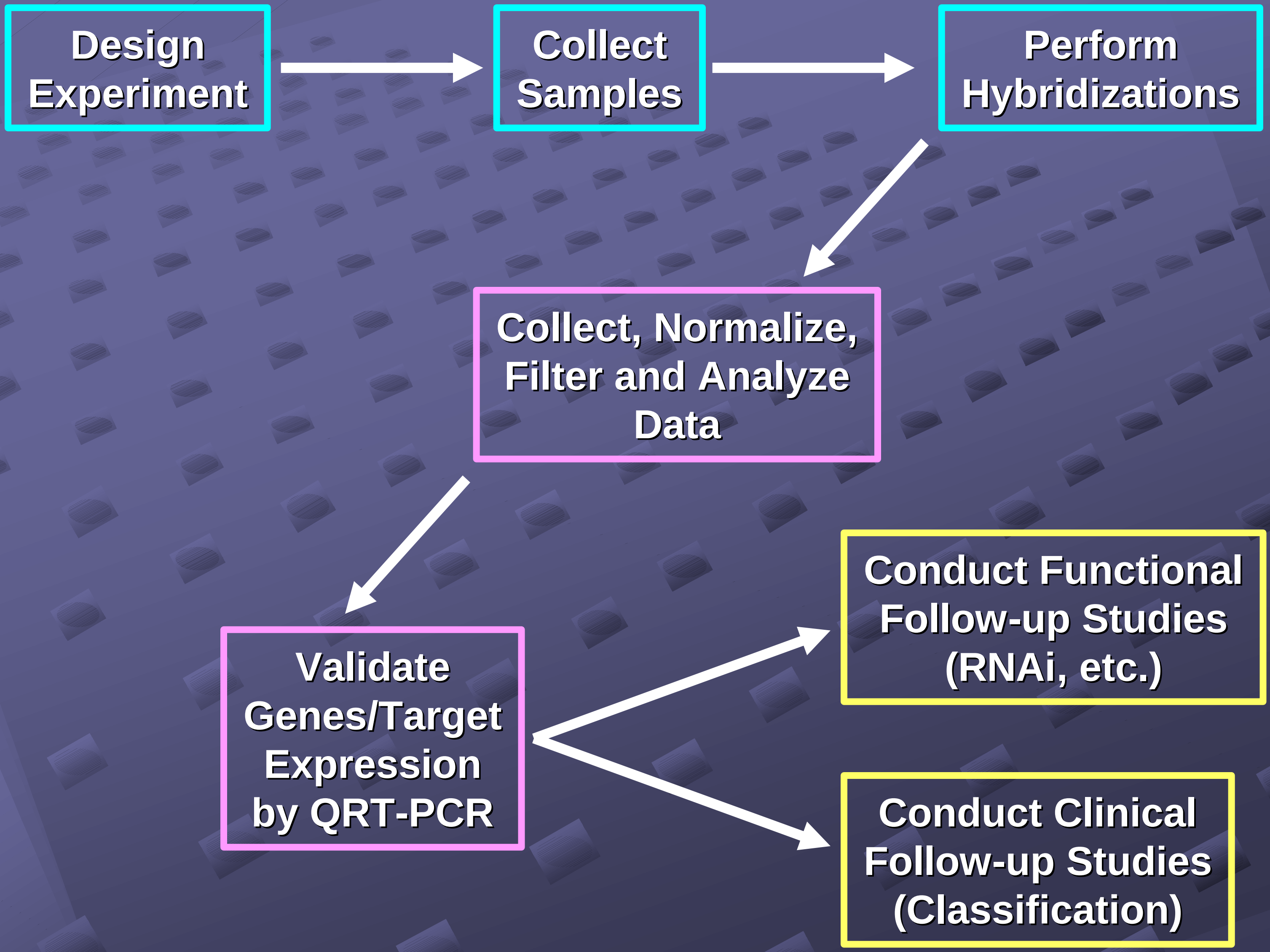
Obtain RNA Samples

Prepare Fluorescently Labeled Targets

PM

MM

# Workflow in an Array Experiment



Design Experiment

Collect Samples

Perform Hybridizations

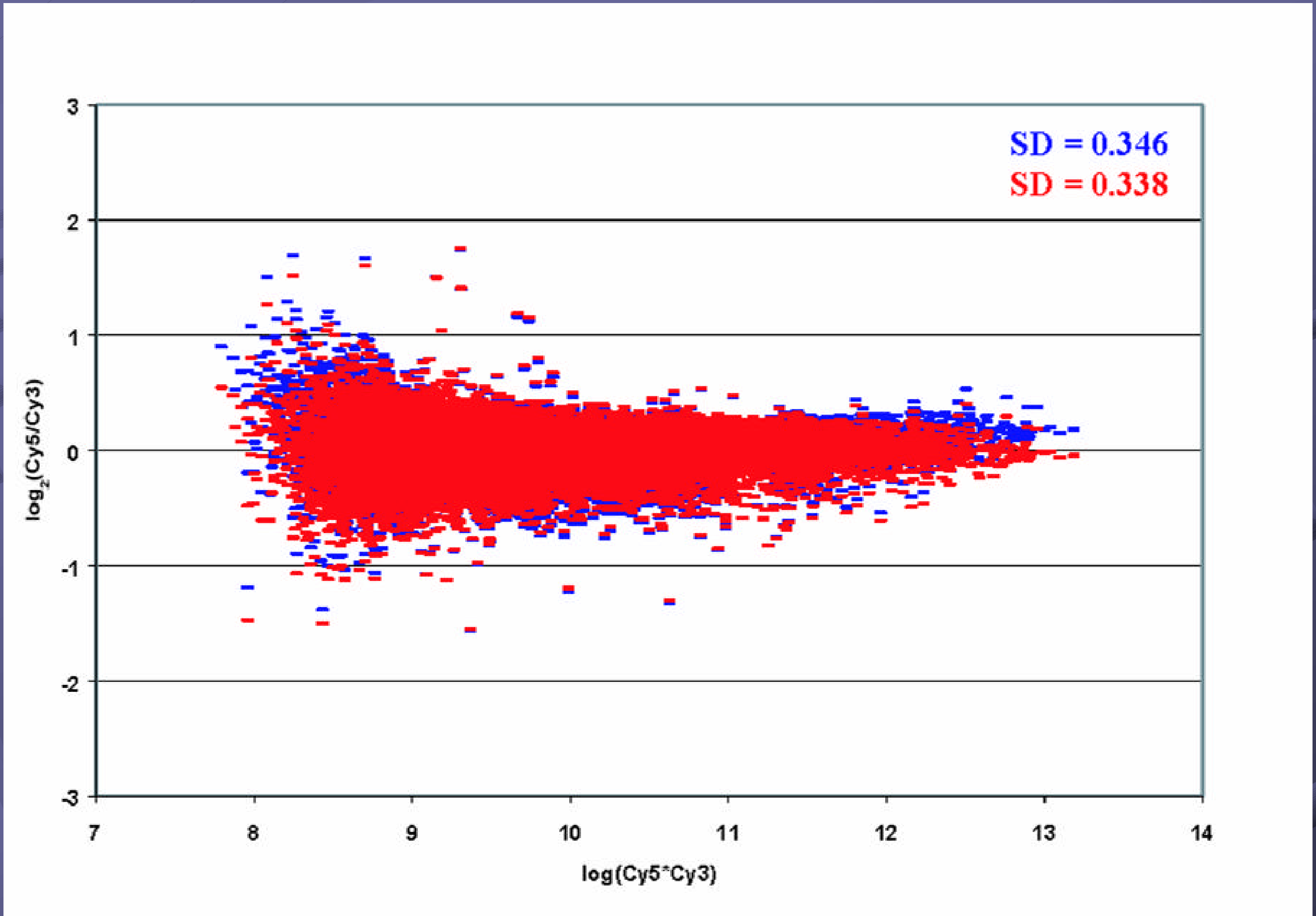
Collect, Normalize, Filter and Analyze Data

Validate Genes/Target Expression by QRT-PCR

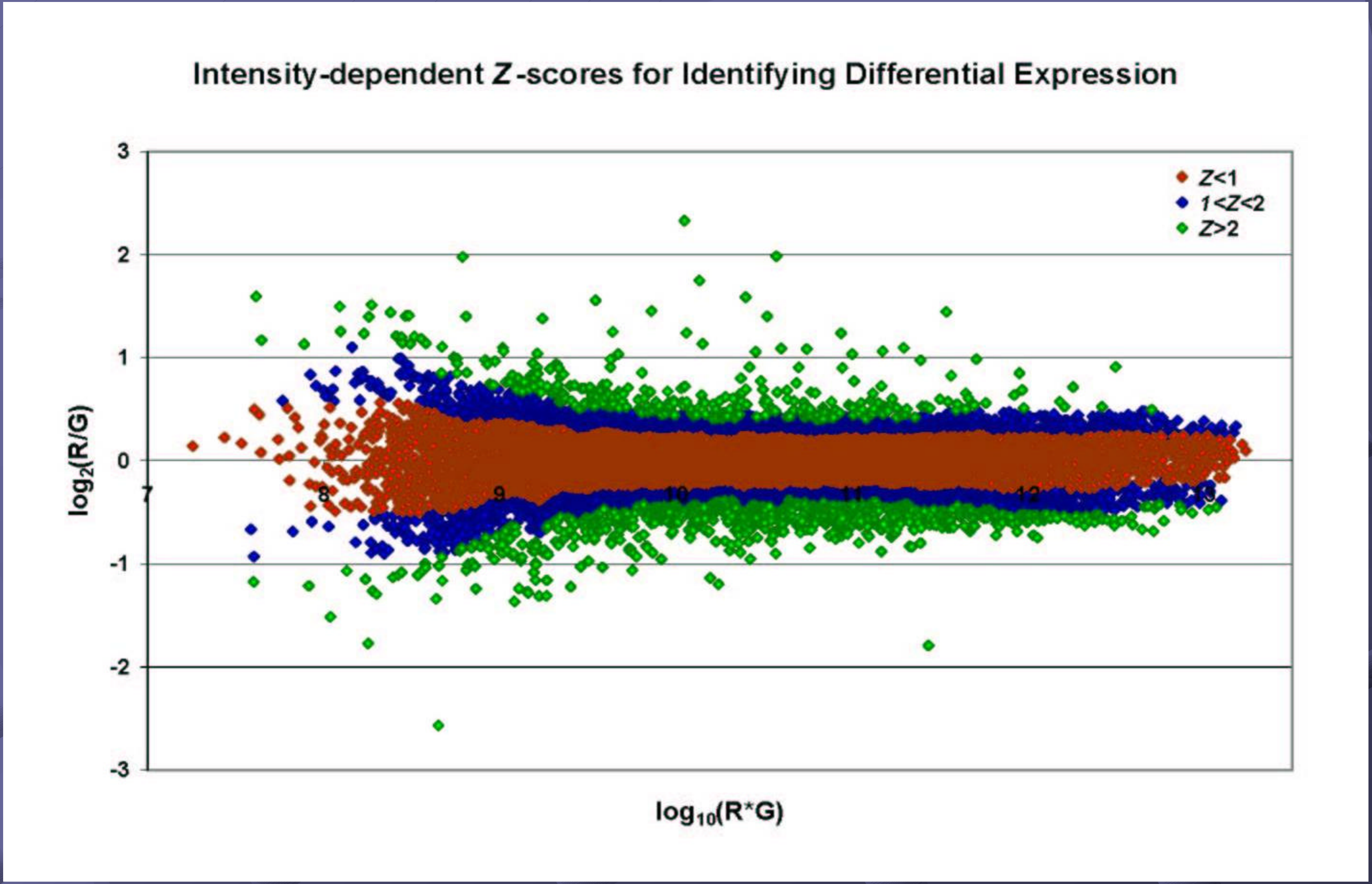
Conduct Functional Follow-up Studies (RNAi, etc.)

Conduct Clinical Follow-up Studies (Classification)

# Why we need a Universal Standard

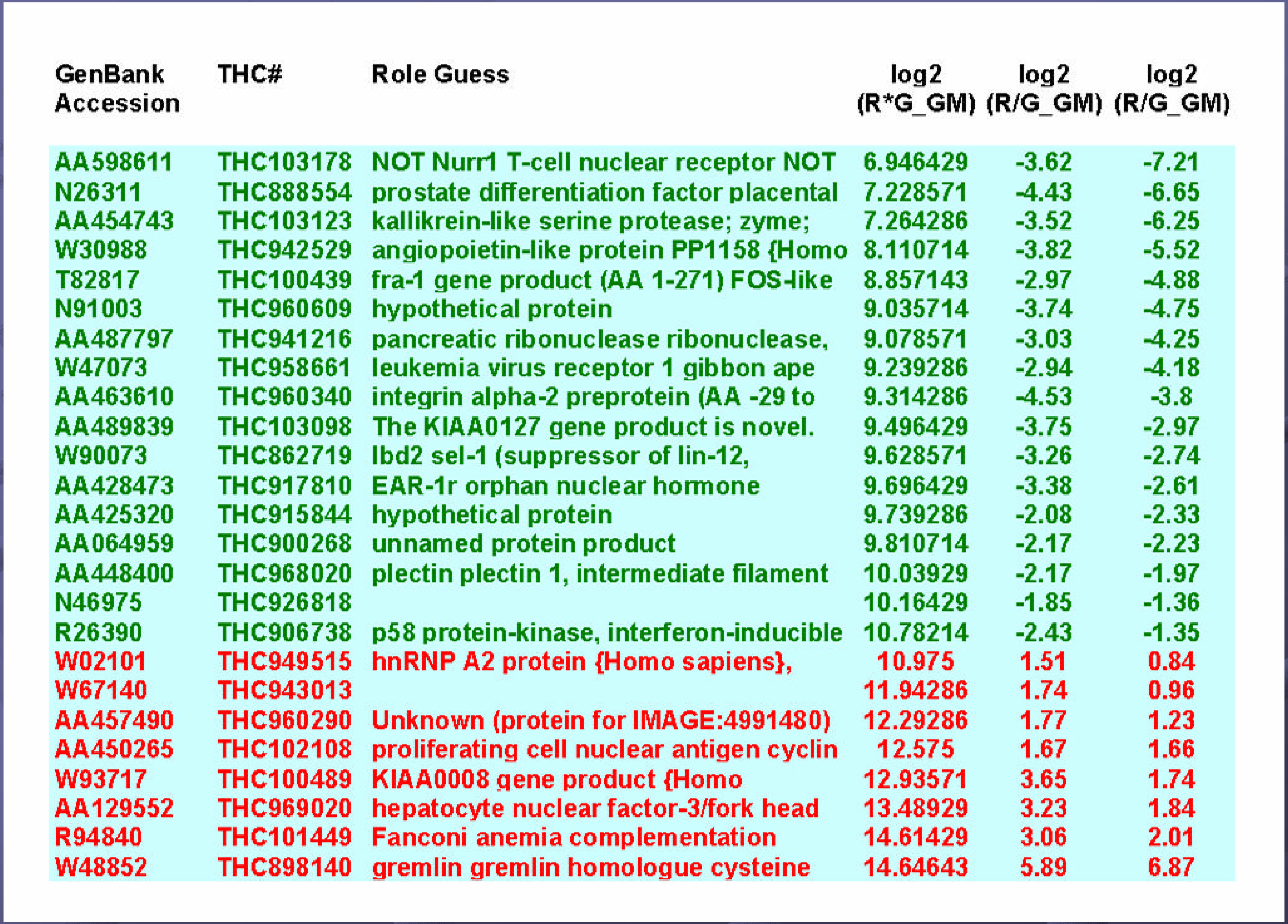


# Intensity-dependent *Z*-score



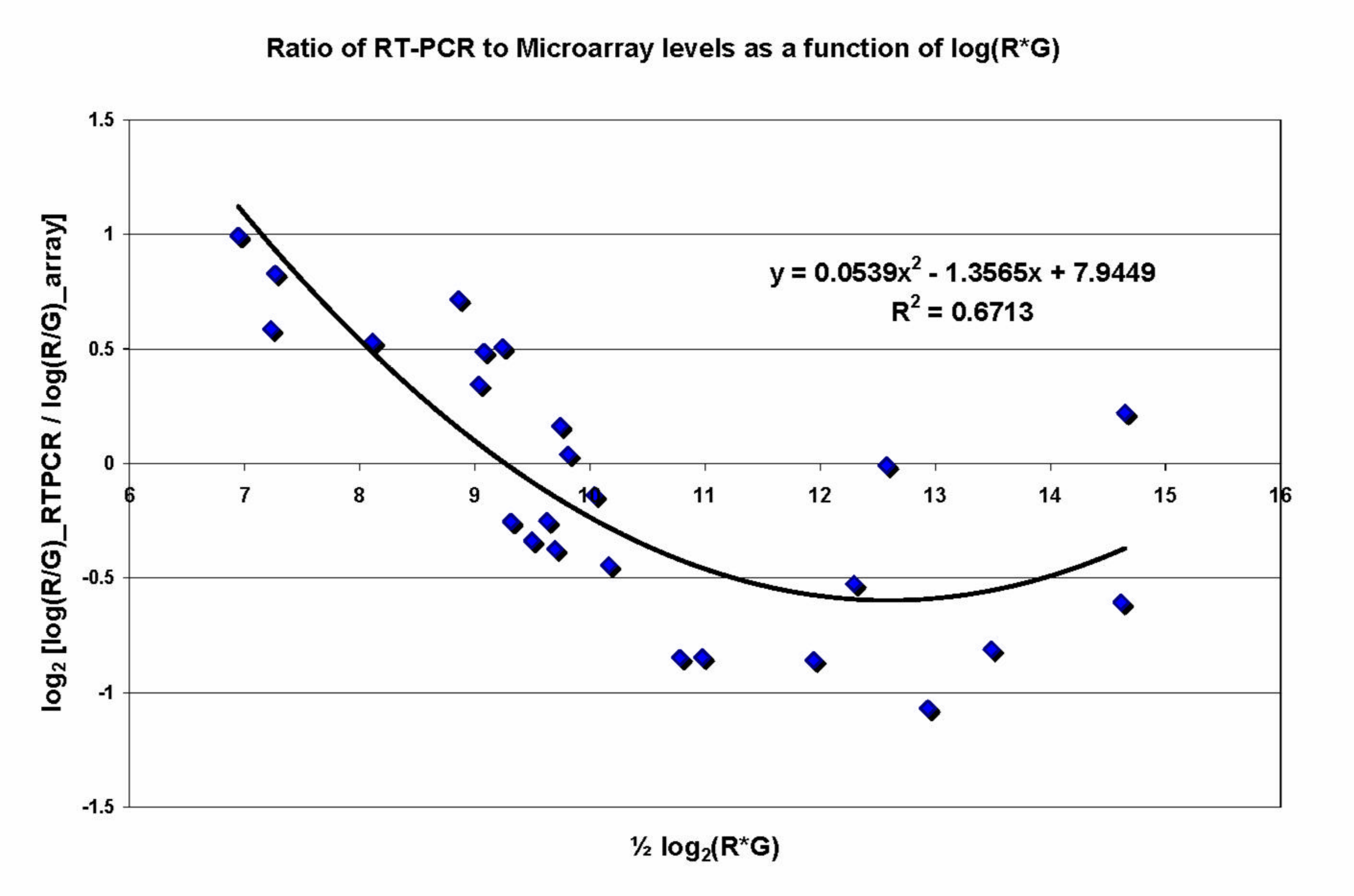
***Z* > 2 is at the 95.5% confidence level**

# Validation



| GenBank Accession | THC# | Role Guess | log2 (R*G_GM) | log2 (R/G_GM) | log2 (R/G_GM) |
|---|---|---|---|---|---|
| AA598611 | THC103178 | NOT Nurr1 T-cell nuclear receptor NOT | 6.946429 | -3.62 | -7.21 |
| N26311 | THC888554 | prostate differentiation factor placental | 7.228571 | -4.43 | -6.65 |
| AA454743 | THC103123 | kallikrein-like serine protease; zyme; | 7.264286 | -3.52 | -6.25 |
| W30988 | THC942529 | angiopoietin-like protein PP1158 {Homo | 8.110714 | -3.82 | -5.52 |
| T82817 | THC100439 | fra-1 gene product (AA 1-271) FOS-like | 8.857143 | -2.97 | -4.88 |
| N91003 | THC960609 | hypothetical protein | 9.035714 | -3.74 | -4.75 |
| AA487797 | THC941216 | pancreatic ribonuclease ribonuclease, | 9.078571 | -3.03 | -4.25 |
| W47073 | THC958661 | leukemia virus receptor 1 gibbon ape | 9.239286 | -2.94 | -4.18 |
| AA463610 | THC960340 | integrin alpha-2 preprotein (AA -29 to | 9.314286 | -4.53 | -3.8 |
| AA489839 | THC103098 | The KIAA0127 gene product is novel. | 9.496429 | -3.75 | -2.97 |
| W90073 | THC862719 | lbd2 sel-1 (suppressor of lin-12, | 9.628571 | -3.26 | -2.74 |
| AA428473 | THC917810 | EAR-1r orphan nuclear hormone | 9.696429 | -3.38 | -2.61 |
| AA425320 | THC915844 | hypothetical protein | 9.739286 | -2.08 | -2.33 |
| AA064959 | THC900268 | unnamed protein product | 9.810714 | -2.17 | -2.23 |
| AA448400 | THC968020 | plectin plectin 1, intermediate filament | 10.03929 | -2.17 | -1.97 |
| N46975 | THC926818 |  | 10.16429 | -1.85 | -1.36 |
| R26390 | THC906738 | p58 protein-kinase, interferon-inducible | 10.78214 | -2.43 | -1.35 |
| W02101 | THC949515 | hnRNP A2 protein {Homo sapiens}, | 10.975 | 1.51 | 0.84 |
| W67140 | THC943013 |  | 11.94286 | 1.74 | 0.96 |
| AA457490 | THC960290 | Unknown (protein for IMAGE:4991480) | 12.29286 | 1.77 | 1.23 |
| AA450265 | THC102108 | proliferating cell nuclear antigen cyclin | 12.575 | 1.67 | 1.66 |
| W93717 | THC100489 | KIAA0008 gene product {Homo | 12.93571 | 3.65 | 1.74 |
| AA129552 | THC969020 | hepatocyte nuclear factor-3/fork head | 13.48929 | 3.23 | 1.84 |
| R94840 | THC101449 | Fanconi anemia complementation | 14.61429 | 3.06 | 2.01 |
| W48852 | THC898140 | gremlin gremlin homologue cysteine | 14.64643 | 5.89 | 6.87 |

# Why we need a Universal Standard



# A Strawman for a Standard

- A set of 1040 synthetic DNAs (“alien DNA”?), at least 500bp in length, cloned into an expression vector (with poly-A), with the clones freely available.
  - Freely available clones
  - A wide range of GC content, etc.
- A set of 70-mers designed using open-source software
- A set of 25-mer/other public probes
- A collection of *in vitro* transcribed RNAs
- A set of standard (minimum 2) mixtures of these RNAs spanning a range of concentrations and fold-changes
  - Concentration range: fewer than “1 transcript per cell” to “1000s per cell” (8 samples for 8 $\log_{10}$s)
  - Fold-change range: 1, ±1.1, ±1.2, ±1.4, ±1.6, ±1.8, ±2, ±4, ±8, ±16, ±32, ±64, on/off (26 points)
  - Multiple samples at each concentration/fold point with range of GC content (5?)
- 1040 = 8*26*5
- Associated Spiking Mixtures for inclusion in real RNA mixtures
- A set of “Standard RNAs” for each species that can serve as a background for these exogenous controls.

# What’s wrong with this picture?

- This *is not* a standard RNA for all comparisons between platforms, patients, labs, nor will it allow particular probes to be validated.
- This approach would allow instrument and software to be validated, but not choices for particular gene probes on the various platforms.
- This may not allow questions to be addressed such as how splice variants and gene families impact expression measurements on a particular platform.
- This will not be limited to a single species.

# The Ultimate Standard?

- 30,000-50,000-100,000 DNAs (one for every gene and variant) cloned into an expression vector (with poly-A), with the clones freely available.
  - Freely available clones
- A set of 70-mers designed using open-source software
- A set of 25-mer/other public probes
- A collection of individual *in vitro* transcribed RNAs
- A set of standard (minimum 2) mixtures of these RNAs spanning a range of concentrations and fold-changes
  - Concentration range: fewer than “1 transcript per cell” to “1000s per cell” (8 samples for 8 $\log_{10}$s)
  - Fold-change range: 1, ±1.1, ±1.2, ±1.4, ±1.6, ±1.8, ±2, ±4, ±8, ±16, ±32, ±64, on/off (26 points)
  - Multiple samples at each concentration/fold point with range of GC content (5?)
- A mixture of RNAs with fixed concentrations for *platform and probe* validation
- Individually-derived RNAs mixtures for each tissue, disease state, developmental stage, ….

# Conclusions

- We need to define common terms to describe our sensitivity, specificity, etc. (in a quantitative way)
- Need to define standards based on particular questions
  - Quality control standards for array platforms to facilitate comparisons and assess dynamic range
  - Quantitative measure of expression for each gene
  - RNA Standards to allow absolute comparisons across datasets
  - Focused standards for a particular experiment